

CLAIMS

Sub
A2

1. A method of preparing a mechanically searchable index with respect to a vector database in which a finite number of sets each including at least N-dimensional real vector and an identification number of the vector are registered as vector data, said method comprising:

a first step of vector index preparation of dividing N components into m sets in a predetermined method with respect to the N-dimensional real vector V of each vector data in said vector database, preparing m partial vectors v_1 to v_m , subsequently tabulating a distribution of a norm of the partial vector v_k ($k = 1$ to m), preparing a norm division table in which a norm range of a predetermined D type norm division is determined, calculating a region number d to which said partial vector v_k belongs in accordance with predetermined D region center vectors p_1 to p_D , tabulating a distribution of a cosine $(v_k \cdot p_d) / (|v_k| * |p_d|)$ of an angle formed by said partial vector v_k and the region center vector p_d as a declination distribution, and preparing a declination division table in which a declination range of the predetermined C type declination division is recorded;

a second step of the vector index preparation of dividing N components into m sets in the same method as said first step with respect to the N-dimensional real vector V of each vector data in said vector database, preparing m partial vectors v_1 to v_m ,

referring to said norm division table to calculate a number r of the norm division to which the norm of said partial vector v_b belongs with respect to the partial vector v_b ($b = 1$ to m) for the partial space number b , calculating the region number d to which

5 said partial vector v_b belongs in accordance with the predetermined D region center vectors p_1 to p_b in the same method as said first step, calculating a declination $(v_b \cdot p_d) / (|v_b| * |p_d|)$ as a cosine of an angle formed by said partial vector v_b and the region center vector p_d indicating a center direction of the region of said region

10 number d , referring to said declination division table, calculating a number c of the belonging declination division, and calculating index registration data to be registered in a vector index from said partial space number b , said region number d , said declination division number c , said norm division number r , the component of

15 said partial vector v_b , and the identification number i ; and

a third step of the vector index preparation of constituting the vector index such that the identification number and the component of each partial vector can be searched using a set of the partial space number b , the region number d , the

20 declination division number c and a norm division number range $[r_1, r_2]$ as a key from said norm division table, said declination division table, and said index registration data, and such that the vector component of each vector data can be searched with the identification number of the vector component.

2. A method of preparing a mechanically searchable index with respect to a vector database in which a finite number of sets each including at least N-dimensional real vector and an identification number of the vector are registered as vector data, said method comprising:

a first step of vector index preparation of dividing N components into m sets in a predetermined method with respect to the N-dimensional real vector V of each vector data in said vector database, preparing m partial vectors v_1 to v_m , subsequently tabulating a distribution of a norm of the partial vector v_b ($b = 1$ to m) for each partial space number b, preparing a norm division table in which a norm range of a predetermined D type norm division is determined, calculating a region number d to which said partial vector v_b belongs in accordance with predetermined D region center vectors p_1 to p_b , tabulating a distribution of a cosine ($(v_b \cdot p_d) / (|v_b| * |p_d|)$) of an angle formed by said partial vector v_b and the region center vector p_d as a declination distribution, and preparing a declination division table in which a declination range of the predetermined C type declination division is recorded;

a second step of the vector index preparation of dividing N components into m sets in the same method as said first step with respect to the N-dimensional real vector V of each vector data in said vector database, preparing m partial vectors v_1 to v_m , referring to said norm division table to calculate a number r of the norm division to which the norm of said partial vector v_b

belongs with respect to the partial vector v_b ($b = 1$ to m) for said partial space b , calculating the region number d to which said partial vector v_b belongs in accordance with the predetermined D region center vectors p_1 to p_b in the same method as said first step, calculating a declination $(v_b \cdot p_d) / (|v_b| * |p_d|)$ as a cosine of an angle formed by said partial vector v_b and the region center vector p_d indicating a center direction of the region of said region number d , referring to said declination division table, calculating a number c of the belonging declination division, calculating a component division number w_j of a predetermined range to which v_{bj} belongs from a maximum value of the norm of the norm division corresponding to said calculated norm division number r with respect to each component v_{bj} of said calculated partial vector v_b , and calculating index registration data to be registered in a vector index from said partial space number b , said region number d , said declination division number c , said norm division number r , a string of said component division numbers w_j , and the identification number i ; and

a third step of the vector index preparation of constituting the vector index such that the identification number and the component of each partial vector can be searched using a set of the partial space number b , the region number d , the declination division number c and a norm division number range $[r_1, r_2]$ as a key from said norm division table, said declination division table, and said index registration data, and such that the vector component of each vector data can be searched with the

identification number of the vector component.

3. The vector index preparing method according to claim
1 or 2 wherein in the first and second steps of said vector index
5 preparation, an angle cosine $(vb \cdot pd) / (|vb| * |pd|)$ is used as a
function of an angle formed by the partial vector vb and the region
center vector pd , and a value of the function is used as a
declination to obtain the declination distribution.

10 4. The vector index preparing method according to claim
1 or 2 wherein in the first and second steps of said vector index
preparation, N/m components or $(N/m)+1$ components are extracted in
order from a top component of V so that all components of an N -
dimensional vector V are extracted, and the partial vector is
15 prepared.

5. The vector index preparing method according to claim
1 wherein in the first step of said vector index preparation,
during preparation of the norm division table, the norm division is
20 determined based on the tabulation result of the norm distribution
so that the number of partial vectors belonging to the norm range
corresponding to each norm division becomes as uniform as possible.

6. The vector index preparing method according to claim
25 1 wherein in the first step of said vector index preparation,

during preparation of the declination division table, the declination division is determined based on the tabulation result of the declination distribution so that the number of partial vectors belonging to the declination range corresponding to each declination division becomes as uniform as possible.

7. The vector index preparing method according to claim 1 or 2 wherein in the first and second steps of said vector index preparation, the region number of the partial vector v_b is obtained as a number d of the region center vector p_d in which a cosine $(v_b \cdot p_d) / (|v_b| * |p_d|)$ of an angle formed by p_d and v_b is largest among the predetermined D region center vector p_1 to p_b .

8. The vector index preparing method according to claim 1 or 2 wherein in the third step of said vector index preparation, a search tree in which a number $(b * Nd * Nc * Nr) + (d * Nc * Nr) + (c * Nr) + r$ obtained by combining the partial space number b , the region number d , the declination division number c , and the norm division number r can be used as a key to search the identification number i and the component of the vector, and a table in which the vector data identification number is used as an affix and the key of said search tree of each partial vector is recorded are prepared and used as part of the vector index.

9. The vector index preparing method according to claim

1 or 2 wherein in the second step of said vector index preparation, the vector obtained by normalizing all vectors $(0, \dots, 0, +1)$ to $(-1, \dots, -1)$ whose component is any one of $\{-1, 0, +1\}$ and which are not 0 vector is used as the region center vector.

5

10. A similar vector searching method in which a query vector Q of an N -dimensional real vector, an inner product lower limit value α , and maximum obtained vector number L are designated as search conditions, a vector index prepared from vector data with a finite number of sets of at least N -dimensional real vector and an ID number of the real vector registered therein is searched, and L sets at maximum $(i, V \cdot Q)$ of an identification number i and an inner product of Q and V are obtained with respect to vector data (i, V) of said vector database whose value $V \cdot Q$ of the inner product with said query vector Q is larger than said inner product lower limit value α , said similar vector searching method comprising:

15 a first step of similar vector search of dividing N components of Q into m sets in the same predetermined method as a method used in preparing said vector index with respect to said query vector Q , preparing m partial query vectors q_1 to q_m ,
20 calculating a partial inner product lower limit value f_b as a lower limit value of an inner product (hereinafter referred to as "partial inner product) of each partial query vector q_b and the corresponding partial vector from a designated inner product lower
25 limit value α , calculating a partial space number b , and a set $(c,$

[r_1, r_2]) of a declination division number c to be searched in a region number d and a norm division range [r_1, r_2] from a value of an inner product $p_d \cdot q_b$ of the region center vector p_d and said partial query vector q_b , said partial inner product lower limit value f_b , and a norm division table and a declination division table in said vector index with respect to each partial query vector q_b ($b = 1$ to m) and each region b , searching a range of said vector index using ($b, d, c, [r_1, r_2]$) as a search condition based on said calculated ($c, [r_1, r_2]$), obtaining the identification number i and the component of the partial vector v_b satisfying the condition as an index search result, calculating a partial inner product difference $(v_b \cdot q_b) - f_b$ as a difference between a partial inner product $v_b \cdot q_b$ of said v_b and q_b and said partial inner product lower limit value f_b , and accumulating (adding) the difference as an inner product difference upper limit value $S[i]$ of the identification number i of an inner product difference table; and

a second step of the similar vector search of searching said vector index with the identification number i in order from a largest value in said inner product difference table $S[i]$ to obtain a vector data component V , calculating an inner product difference value $t = V \cdot Q - \alpha$ by subtracting α from the inner product $V \cdot Q$ of V and said query vector Q , and outputting a set of at least the identification number i and an inner product $t + \alpha$ as a search result with respect to L pieces at maximum of vector data with a large inner product difference value when L or more pieces of

vector data having the inner product difference value larger than a maximum value of an element having a non-calculated inner product difference value are collected, or when the inner products of all the vector data having a positive inner product difference upper limit value are calculated in said inner product difference table.

11. A similar vector searching method in which a query vector Q of an N -dimensional real vector, a distance upper limit value α , and maximum obtained vector number L are designated as search conditions, a vector index prepared from vector data with a finite number of sets of at least N -dimensional real vector and an identification number of the real vector registered therein is searched, and L sets at maximum (i, p) of an identification number i of an N -dimensional real vector V in said vector data and a distance p between Q and V are obtained such that a value of an inner product with said query vector Q is not more than said distance upper limit value α , said similar vector searching method comprising:

a first step of similar vector search of dividing N components of Q into m sets in the same predetermined method as a method used in preparing said vector index with respect to said query vector Q , preparing m partial query vectors q_1 to q_m , calculating a partial square distance upper limit value f_b as an upper limit value of a square distance $|v_b - q_b|^2$ (i.e., square of Euclidean distance, hereinafter referred to as "partial square

distance") of each partial query vector q_b and the corresponding partial vector v_b from a designated distance upper limit value α , systematically generating a set $(b, d, c, [r_1, r_2])$ of a partial space number b to be searched, a region number d , a declination division number c and a norm division range $[r_1, r_2]$ from said partial query vector q_b , said partial square distance upper limit value f_b , and a norm division table and a declination division table in said vector index with respect to each partial query vector q_b ($b = 1$ to m), searching a range of said vector index using said generated $(b, d, c, [r_1, r_2])$ as a search condition, obtaining the identification number i and the component of the partial vector v_b satisfying the condition as an index search result, calculating a partial square distance difference $f_b - |v_b - q_b|^2$ as a difference between said partial square distance upper limit value f_b and a partial square distance $|v_b - q_b|^2$ of v_b and q_b , and accumulating (adding) the difference as a square distance difference upper limit value $S[i]$ of the identification number i of a square distance difference table; and

a second step of the similar vector search of searching said vector index with the identification number i in order from a largest value in said square distance difference table $S[i]$ to obtain a vector data component V , calculating a square distance difference value $\alpha^2 - |V - Q|^2$ by subtracting a square distance $|V - Q|^2$ of V and said query vector Q from a squared distance upper limit value α^2 , and outputting a set of at least the identification

number i and a distance $(\alpha^2 - t)^{1/2}$ as a search result with respect to L pieces at maximum of vector data with a large square distance difference value t when L or more pieces of vector data having the square distance difference value larger than a maximum value of an element having a non-calculated square distance difference value are collected, or when the square distance difference values of all the vector data having a positive square distance difference upper limit value are calculated in said square distance difference table.

12. The similar vector searching method according to claim 10 or 11 wherein in the first step of said similar vector search, N/m components or $(N/m)+1$ components are extracted in order from a top component of V so that all components of an N -dimensional vector V are extracted, and the partial query vector is prepared.

13. The similar vector searching method according to claim 11 wherein in the first step of said similar vector search, the partial inner product lower limit value f_b as the lower limit value of the inner product of said partial query vector q_b and the corresponding partial vector v_b is calculated from a designated inner product lower limit value α by $f_b = \alpha |q_b|^2 / \Sigma(|q_b|^2)$.

14. The similar vector searching method according to claim 11 wherein in the first step of said similar vector search,

the partial square distance upper limit value f_b as the upper limit value of the square distance of said partial query vector q_b and the corresponding partial vector v_b is calculated from a designated distance lower/upper limit value α by $f_b = \alpha^2 |q_b|^2 / \Sigma (|q_b|^2)$.

5

15. An apparatus for preparing a mechanically searchable index with respect to a vector database in which a finite number of sets each including at least N-dimensional real vector and an identification number of the vector are registered as vector data, said apparatus comprising:

10

partial vector calculation means for dividing N components into m sets in a predetermined method with respect to the N-dimensional real vector V of each vector data in said vector database, and preparing m partial vectors v_1 to v_m ;

15

norm distribution tabulation means for tabulating a distribution of a norm of the partial vector v_k ($k = 1$ to m) among said prepared m partial vectors v_1 to v_m , and preparing a norm division table in which a norm range of a predetermined D type norm division is determined;

20

region number calculation means for calculating a region number d to which said partial vector v_k belongs in accordance with predetermined D region center vectors p_1 to p_b ;

25

declination distribution tabulation means for tabulating a distribution of a cosine $(v_k \cdot p_d) / (|v_k| * |p_d|)$ of an angle formed by said partial vector v_k and the region center vector p_d as a

declination distribution, and preparing a declination division table in which a declination range of the predetermined C type declination division is recorded;

5 norm division number calculation means for referring to said norm division table to calculate a number r of the norm division to which the norm of said partial vector v_b belongs with respect to the partial vector v_b ($b = 1$ to m) for the partial space number b among the m partial vectors v_1 to v_m prepared by said partial vector calculation means;

10 declination division number calculation means for calculating a declination $(v_b \cdot p_d) / (|v_b| \cdot |p_d|)$ as a cosine of an angle formed by said partial vector v_b and the region center vector p_d indicating a center direction of the region of said region number d calculated by said region number calculation means;

15 index data calculation means for calculating index registration data to be registered in a vector index from said partial space number b , said region number d , said declination division number c , said norm division number r , the component of said partial vector v_b , and the identification number i ; and

20 index constituting means for constituting the vector index such that the identification number and the component of each partial vector can be searched using a set of the partial space number b , the region number d , the declination division number c and a norm division number range $[r_1, r_2]$ as a key from said norm division table, said declination division table, and said index

25

registration data, and such that the vector component of each vector data can be searched with the identification number of the vector component.

5 16. An apparatus for preparing a mechanically searchable index with respect to a vector database in which a finite number of sets each including at least N-dimensional real vector and an identification number of the vector are registered as vector data, said apparatus comprising:

10 partial vector calculation means for dividing N components into m sets in a predetermined method with respect to the N-dimensional real vector V of each vector data in said vector database, and preparing m partial vectors v_1 to v_m ;

15 norm distribution tabulation means for tabulating a distribution of a norm of the partial vector v_b ($b = 1$ to m) for a partial space number b among said prepared m partial vectors v_1 to v_m , and preparing a norm division table in which a norm range of a predetermined D type norm division is determined;

20 region number calculation means for calculating a region number d to which said partial vector v_b belongs in accordance with predetermined D region center vectors p_1 to p_b ;

25 declination distribution tabulation means for tabulating a distribution of a cosine $(v_b \cdot p_d) / (|v_b| * |p_d|)$ of an angle formed by said partial vector v_b and the region center vector p_d as a declination distribution, and preparing a declination division

table in which a declination range of the predetermined C type declination division is recorded;

norm division number calculation means for referring to said norm division table to calculate a number r of the norm division to which the norm of said partial vector v_b belongs with respect to the partial vector v_b ($b = 1$ to m) for a partial space b among the m partial vectors v_1 to v_m prepared by said partial vector calculation means;

declination division number calculation means for calculating a declination $(v_b \cdot p_d) / (|v_b| * |p_d|)$ as a cosine of an angle formed by said partial vector v_b and the region center vector p_d indicating a center direction of the region of the region number d calculated by said region number calculation means;

component division number calculation means for calculating a component division number w_j of a predetermined range to which v_{bj} belongs from a maximum value of the norm of the norm division corresponding to said calculated norm division number r with respect to each component v_{bj} of said calculated partial vector v_b ;

index data calculation means for calculating index registration data to be registered in a vector index from said partial space number b, said region number d, said declination division number c, said norm division number r, a string of said component division numbers w_j , and the identification number i; and

index constituting means for constituting the vector

index such that the identification number and the component of each partial vector can be searched using a set of the partial space number b, the region number d, the declination division number c and a norm division number range $[r_1, r_2]$ as a key from said norm division table, said declination division table, and said index registration data, and such that the vector component of each vector data can be searched with the identification number of the vector component.

10 17. The vector index preparing apparatus according to claim 15 or 16 wherein said partial vector calculation means extracts N/m components or $(N/m)+1$ components in order from a top component of V so that all components of an N -dimensional vector V are extracted, and prepares the partial vector.

15 18. The vector index preparing apparatus according to claim 15 wherein during preparation of the norm division table said norm distribution tabulation means determines the norm division based on the tabulation result of the norm distribution so that the number of partial vectors belonging to the norm range corresponding to each norm division becomes as uniform as possible.

20 19. The vector index preparing apparatus according to claim 15 wherein during preparation of the declination division table, said declination distribution tabulation means determines
25

the declination division based on the tabulation result of the declination distribution so that the number of partial vectors belonging to the declination range corresponding to each declination division becomes as uniform as possible.

5

20. The vector index preparing apparatus according to claim 15 or 16 wherein said region number calculation means obtains the region number of the partial vector v_b as a number d of the region center vector p_d in which a cosine $(v_b \cdot p_d) / (|v_b| * |p_d|)$ of an angle formed by p_d and v_b is largest among the predetermined D region center vector p_1 to p_D .

10

21. The vector index preparing apparatus according to claim 15 or 16 wherein said index constituting means prepares a search tree in which a number $(b * Nd * Nc * Nr) + (d * Nc * Nr) + (c * Nr) + r$ obtained by combining the partial space number b , the region number d , the declination division number c , and the norm division number r can be used as a key to search the identification number i and the component of the vector, and a table in which the vector data identification number is used as an affix and the key of said search tree of each partial vector is recorded, and uses the search tree and the table as a part of the vector index.

15

20

22. The vector index preparing apparatus according to claim 15 or 16 wherein said region number calculation means uses

25

the vector obtained by normalizing all vectors $(0, \dots, 0, +1)$ to $(-1, \dots, -1)$ whose component is any one of $\{-1, 0, +1\}$ and which are not 0 vector as the region center vector.

5 23. A similar vector searching apparatus for designating
a query vector Q of an N -dimensional real vector, an inner product
lower limit value α , and maximum obtained vector number L as
search conditions, searching a vector index prepared from vector
10 vector and an ID number of the real vector registered therein, and
obtaining L sets at maximum $(i, V \cdot Q)$ of an identification number i
and an inner product of Q and V with respect to vector data (i, V)
of said vector database whose value $V \cdot Q$ of the inner product with
said query vector Q is larger than said inner product lower limit
15 value α , said similar vector searching apparatus comprising:

partial query condition calculation means for dividing N
components of Q into m sets in the same predetermined method as a
method used in preparing said vector index with respect to said
query vector Q , preparing m partial query vectors q_1 to q_m , and
20 calculating a partial inner product lower limit value f_b as a lower
limit value of an inner product (hereinafter referred to as
"partial inner product") of each partial query vector q_b and the
corresponding partial vector from a designated inner product lower
limit value α ;

25 search object range generation means for calculating a

partial space number b , and a set $(c, [r_1, r_2])$ of a declination division number c to be searched in a region number d and a norm division range $[r_1, r_2]$ from a value of an inner product $p_d \cdot q_b$ of the region center vector p_d and said partial query vector q_b , said
5 partial inner product lower limit value f_b , and a norm division table and a declination division table in said vector index with respect to each partial query vector q_b ($b = 1$ to m) and each region b ;

index search means for searching a range of said vector
10 index using $(b, d, c, [r_1, r_2])$ as a search condition based on $(c, [r_1, r_2])$ calculated by said search object range generation means, and obtaining the identification number i and the component of the partial vector v_b satisfying the condition as an index search result;

15 inner product difference upper limit calculation means for calculating a partial inner product difference $(v_b \cdot q_b) - f_b$ as a difference between a partial inner product $v_b \cdot q_b$ of said v_b and q_b and said partial inner product lower limit value f_b , and accumulating (adding) the difference as an inner product difference
20 upper limit value $S[i]$ of the identification number i of an inner product difference table; and

similarity search result determination means for searching said vector index with the identification number i in order from a largest value in said inner product difference table
25 $S[i]$ to obtain a vector data component V , calculating an inner

product difference value $t = V \cdot Q - \alpha$ by subtracting α from the inner product $V \cdot Q$ of V and said query vector Q , and outputting a set of at least the identification number i and an inner product $t + \alpha$ as a search result with respect to L pieces at maximum of vector data with a large inner product difference value when L or more pieces of vector data having the inner product difference value larger than a maximum value of an element having a non-calculated inner product difference value are collected, or when the inner products of all the vector data having a positive inner product difference upper limit value are calculated in said inner product difference table.

24. A similar vector searching apparatus for designating a query vector Q of an N -dimensional real vector, a distance upper limit value α , and maximum obtained vector number L as search conditions, searching a vector index prepared from vector data with a finite number of sets of at least N -dimensional real vector and an identification number of the real vector registered therein, and obtaining L sets at maximum (i, p) of an identification number i of an N -dimensional real vector V in said vector data and a distance p between Q and V such that a value of an inner product with said query vector Q is not more than said distance upper limit value α , said similar vector searching apparatus comprising:

partial query condition calculation means for dividing N components of Q into m sets in the same predetermined method as a

method used in preparing said vector index with respect to said query vector Q , preparing m partial query vectors q_1 to q_m , calculating a partial square distance upper limit value f_b as an upper limit value of a square distance $|v_b - q_b|^2$ (i.e., square of Euclidean distance, hereinafter referred to as "partial square distance") of each partial query vector q_b and the corresponding partial vector v_b from a designated distance upper limit value α ;

search object range generation means for systematically generating a set $(b, d, c, [r_1, r_2])$ of a partial space number b to be searched, a region number d , a declination division number c and a norm division range $[r_1, r_2]$ from said partial query vector q_b , said partial square distance upper limit value f_b , and a norm division table and a declination division table in said vector index with respect to said partial query vector q_b ($b = 1$ to m);

index search means for searching a range of said vector index using $(b, d, c, [r_1, r_2])$ generated by said search object range generation means as a search condition, and obtaining the identification number i and the component of the partial vector v_b satisfying the condition as an index search result;

square distance difference upper limit calculation means for calculating a partial square distance difference $f_b - |v_b - q_b|^2$ as a difference between said partial square distance upper limit value f_b and a partial square distance $|v_b - q_b|^2$ of v_b and q_b , and accumulating (adding) the difference as a square distance difference upper limit value $S[i]$ of the identification number i of

a square distance difference table; and

similarity search result determination means for searching said vector index with the identification number i in order from a largest value in said square distance difference table $S[i]$ to obtain a vector data component V , calculating a square distance difference value $\alpha^2 - |V-Q|^2$ by subtracting a square distance $|V-Q|^2$ of V and said query vector Q from a squared distance upper limit value α^2 , and outputting a set of at least the identification number i and a distance $(\alpha^2 - t)^{1/2}$ as a search result with respect to L pieces at maximum of vector data with a large square distance difference value t when L or more pieces of vector data having the square distance difference value larger than a maximum value of an element having a non-calculated square distance difference value are collected, or when the square distance difference values of all the vector data having a positive square distance difference upper limit value are calculated in said square distance difference table.

25. The similar vector searching apparatus according to claim 23 or 24 wherein said partial query condition calculation means extracts N/m components or $(N/m)+1$ components in order from a top component of V so that all components of an N -dimensional vector V are extracted, and prepares the partial query vector.

26. The similar vector searching apparatus according to

claim 23 wherein the partial inner product lower limit value f_b as the lower limit value of the inner product of said partial query vector q_b and the corresponding partial vector v_b is calculated from a designated inner product lower limit value α by $f_b =$

5 $\alpha |q_b|^2 / \Sigma (|q_b|^2).$

27. The similar vector searching apparatus according to claim 24 wherein the partial square distance upper limit value f_b as the upper limit value of the square distance of said partial query vector q_b and the corresponding partial vector v_b is
10 calculated from a designated distance lower/upper limit value α by $f_b = \alpha^2 |q_b|^2 / \Sigma (|q_b|^2).$

28. A recording medium in which a computer program for
15 executing the method of claim 1 or 2 is recorded.

29. A recording medium in which a computer program for realizing the apparatus of claim 15 or 16 by software is recorded.